Homework Assignment 2

Part1: Census Data

- ***Please note: Full script is included at the end of this document***

- *Data Exploration*
  - The CSV did indeed contain a large amount of NA values, which were indicated by a question mark. These had to be designated NA properly, which also included removing the trailing white space.

```
censusData <- read.csv("adult.csv", strip.white=TRUE)
censusData$workclass <- gsub("?", NA, censusData$workclass, fixed=TRUE)
censusData$workclass <- as.factor(censusData$workclass)
```

  - The summary statistics show that, interestingly enough, many of the variables in the dataset were normally distributed with very little skew. Besides viewing these variables in histogram and plot form, this was also obvious because of the low kurtosis values (<1, in many cases) for many of the variables.

```
                vars     n      mean        sd median   trimmed      mad   min      max    range  skew kurtosis      se
age                1 32561     38.58     13.64     37     37.69    14.83    17       90       73  0.56    -0.17    0.08
workclass*         2 30725      4.10      1.14      4      4.11     0.00     1        8        7  0.06     1.98    0.01
fnlwgt             3 32561 189778.37 105549.98 178356 180802.36 88798.84 12285  1484705  1472420  1.45     6.22  584.94
education*         4 32561     11.30      3.87     12     11.81     2.97     1       16       15 -0.93     0.68    0.02
education.number   5 32561     10.08      2.57     10     10.19     1.48     1       16       15 -0.31     0.62    0.01
marital.status*    6 32561      3.61      1.51      3      3.65     2.97     1        7        6 -0.01    -0.54    0.01
occupation*        7 32561      7.57      4.23      8      7.50     5.93     1       15       14  0.11    -1.23    0.02
race*              8 32561      4.67      0.85      5      4.90     0.00     1        5        4 -2.44     4.87    0.00
sex*               9 32561      1.67      0.47      2      1.71     0.00     1        2        1 -0.72    -1.48    0.00
capital.gain      10 32561   1077.65   7385.29      0      0.00     0.00     0    99999    99999 11.95   154.77   40.93
capital.loss      11 32561     87.30    402.96      0      0.00     0.00     0     4356     4356  4.59    20.37    2.23
hours.per.week    12 32561     40.44     12.35     40     40.55     4.45     1       99       98  0.23     2.92    0.07
native.country*   13 32561     37.72      7.82     40     40.00     0.00     1       42       41 -3.66    12.53    0.04
salary*           14 32561      1.24      0.43      1      1.18     0.00     1        2        1  1.21    -0.53    0.00
income            15 32561      0.24      0.43      0      0.18     0.00     0        1        1  1.21    -0.53    0.00
relationship      16 32561      0.47      0.50      0      0.47     0.00     0        1        1  0.11    -1.99    0.00
newSex            17 32561      0.33      0.47      0      0.29     0.00     0        1        1  0.72    -1.48    0.00
```

  - I also created a correlation matrix for many of the variables in the DF. The strongest correlations were related to income, with age (.23), education level (.33), hours per week (.22) and sex (.22) representing some of the higher scores. The highest, however, was the correlation between marital status and income (.43).

```
> corrDF <- subset(censusData, select=c("age", "education.number", "hours.per.week", "income", "capital.gain", "capital.loss", "newSex", "relationship"))
> cor(corrDF)
                      age education.number hours.per.week     income capital.gain capital.loss      newSex relationship
age            1.00000000       0.03652719     0.06875571  0.2340371   0.07767450   0.05777454 -0.08883173   0.31823901
education.number 0.03652719     1.00000000     0.14812273  0.3351540   0.12263011   0.07992296 -0.01228005   0.07825754
hours.per.week 0.06875571       0.14812273     1.00000000  0.2296891   0.07840862   0.05425636 -0.22930915   0.21091226
income         0.23403710       0.33515395     0.22968907  1.0000000   0.22332882   0.15052631 -0.21598015   0.43494363
capital.gain   0.07767450       0.12263011     0.07840862  0.2233288   1.00000000  -0.03161506 -0.04847965   0.08411882
capital.loss   0.05777454       0.07992296     0.05425636  0.1505263  -0.03161506   1.00000000 -0.04556735   0.07813040
newSex        -0.08883173      -0.01228005    -0.22930915 -0.2159802  -0.04847965  -0.04556735  1.00000000  -0.42146462
relationship   0.31823901       0.07825754     0.21091226  0.4349436   0.08411882   0.07813040 -0.42146462   1.00000000
```

- *Teach me something*
  - This type of demographic data provides useful information for marketers and product designers. In the case of product, these correlations provide an empirical starting point to justifying the creation of digital products for educated, married adults. Since this subset of the population is significantly correlated with higher incomes, they are probably also more likely to be willing to spend money on in-app purchases and digital subscriptions, which have been shown to be considerably elastic for lower earners.

Since most digital products are designed for teenagers and young adults, I think there is a defensible business case in developing targeted and niche software (and hardware) that appeal to these more affluent demographic subsets. Simply put, we need to make stuff for people with money and these data tell us who they are.

- *Important considerations*
  - Besides the obvious considerations, such as the fact that correlations do not indicate causation and can often be very misleading, other important points need to be made about this dataset. First, while some of the variables are indeed normally distributed and representative, not all are. Notably, gender is highly biased, with males representing around two-thirds of the dataset.
  - Also concerning is capital gains/losses. While this sounds like it could have been a very useful indicator of wealth, I chose not to consider it in my analysis. This is because capital gains and losses are only those that are realized, and thus taxable. A large amount of wealth is tied up in long-term investment vehicles that are not subject to frequent redemptions, and thus not represented in this dataset. Furthermore, any meaningful analysis with this variable would probably have to include the absolute value of gains AND losses, as, unsurprisingly, both are *positively* correlated with income.
- Run a logistic regression
  - I ran a logistic regression using the variables I suspected were most useful for predicting if income would be above or below $50,000. At first, I was concerned about multicolinearity, given the moderate degree of correlation between the IVs in the data. However, as shown below, the variance intensity factor (VIF) was well under 10, which is considered a reasonable threshold.

```
glm(formula = income ~ education.number + age + relationship +
    capitalTotal + hours.per.week + newSex, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2938  -0.5554  -0.2407  -0.0525   3.3270

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -9.353e+00  1.625e-01 -57.565  < 2e-16 ***
education.number  3.653e-01  9.130e-03  40.008  < 2e-16 ***
age               3.055e-02  1.735e-03  17.608  < 2e-16 ***
relationship      2.329e+00  5.391e-02  43.194  < 2e-16 ***
capitalTotal      3.420e-04  1.254e-05  27.271  < 2e-16 ***
hours.per.week    2.950e-02  1.813e-03  16.275  < 2e-16 ***
newSex            1.729e-01  5.608e-02   3.083  0.00205 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24884  on 22184  degrees of freedom
Residual deviance: 15403  on 22178  degrees of freedom
AIC: 15417

Number of Fisher Scoring iterations: 7
```

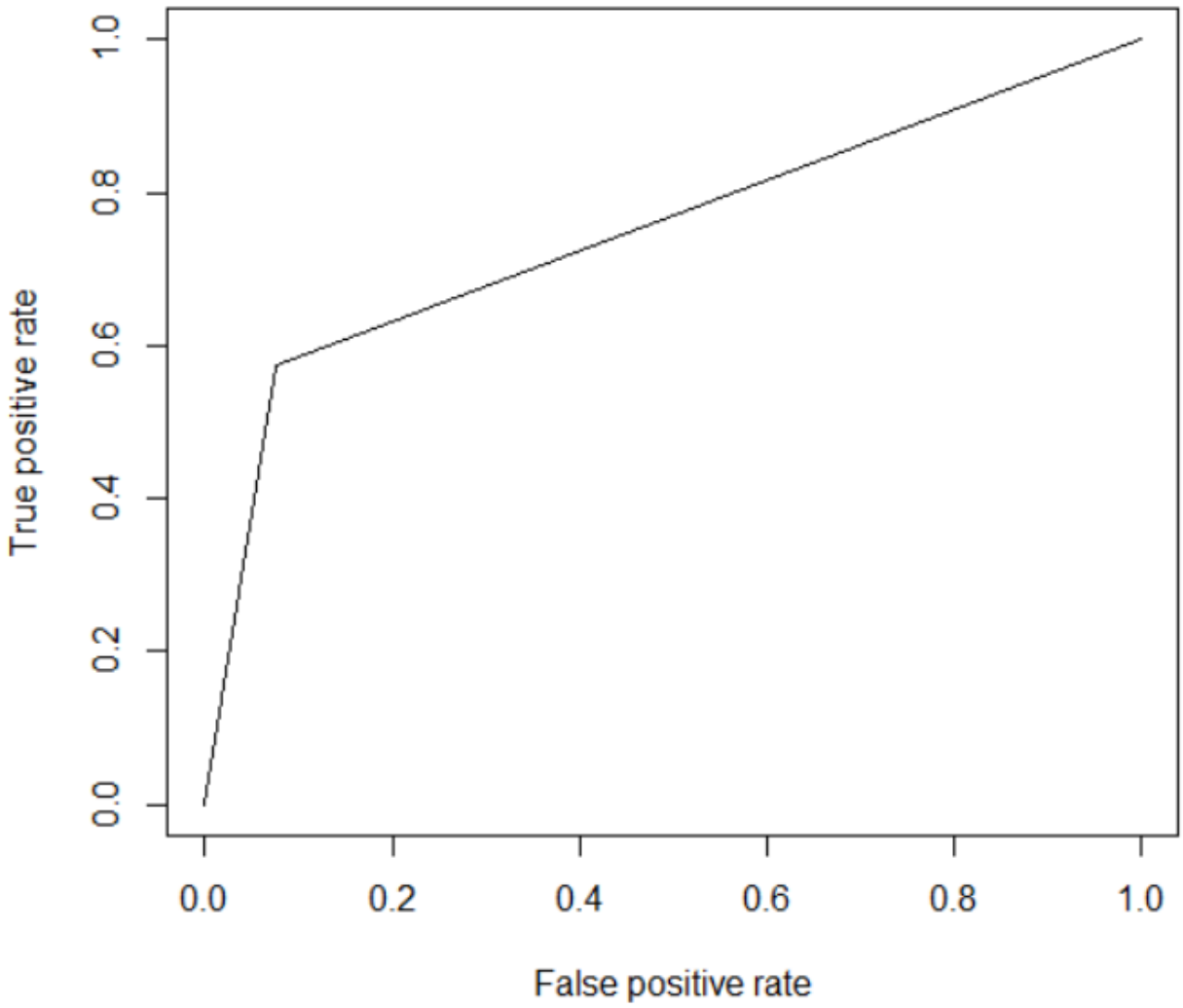
```
> vif(logitmodel)
education.number              age     relationship     capitalTota
        1.051942         1.031263         1.236619         1.01072
  hours.per.week           newSex
        1.040075         1.215153
```

- To check the predicative ability and overall validity of the model, I split the data into a training and testing subset. The overall accuracy of the dataset was calculated to be a considerably high 83.6%, while McFadden's "pseudo" r-squared was .57, which is actually quite high, given that this statistic is not as generous as normal r-squared.

```
> fitted.results <- predict(logitmodel, newdata=subset(test,select=c("age",
urs.per.week", "education.number", "income", "newSex", "relationship", "cap
Total")), type='response')
> fitted.results <- ifelse(fitted.results > 0.5, 1, 0)
> misClasificError <- mean(fitted.results != test$income)
> print(paste('Accuracy', 1-misClasificError))
[1] "Accuracy 0.836299765807963"
```

```
> nullmod <- glm(censusData$income~1, family="binomial")
> mcfaddenR2 <- 1-logLik(logitmodel)/logLik(nullmod)
> mcfaddenR2
'log Lik.' 0.5715237 (df=7)
```

- I also displayed the results of the prediction in the form of a confusion matrix and an ROC curve. For the ROC curve, the area under the curve (AUC) was not as high as I anticipated (.75), which was disappointing but still quite high given the data.



o

```
> confusionMatrix(fitted.results, test$income)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5916  908
         1  490 1226
```

o

```
> auc <- performance(pr, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
[1] 0.7490086
```

Part 2: Progresso Soup

- ***Please note: Full script is included at the end of this document***

- *Create a dummy variable for "Winter" months defined as Oct, Nov, Dec, Jan & Feb. Use the "Month" variable to create this.*

```
soup["Season"] <- NULL
for (i in 1:nrow(soup)) {
  if (soup$Month[i] >= 10 || soup$Month[i] <=2) {
    soup$Season[i] <- 1
  }
  else {
    soup$Season[i] <- 0
  }
}
```

- *Compute the "Market Share" for Progresso (as percentage of total sales) in the Winter vs. non-Winter months using the variable created in (1).*

```
> shareWinter <- (sum(soup$Sales.Progresso[soup$Season == 1]) / sum(so
up$Category_Sales[soup$Season == 1])) * 100
> shareNonWinter <- (sum(soup$Sales.Progresso[soup$Season == 0]) / sum
(soup$Category_Sales[soup$Season == 0])) * 100
> shareNonWinter
[1] 19.92817
> shareWinter
[1] 28.46215
```

- *Develop a linear regression model to predict Progresso sales*
  - *The following is the R code I used to create the model:*

```
lm(formula = soup$Sales.Progresso ~ Price.Campbell + Price.PL +
    Price.Progresso + South + West + East + High_Income + Low_Income +
    Season, data = soup)
```

```
Coefficients:
                 Estimate Std. Error  t value             Pr(>|t|)
(Intercept)       3086.50      48.94   63.070 < 0.0000000000000002 ***
Price.Campbell     915.70      30.49   30.033 < 0.0000000000000002 ***
Price.PL           613.87      32.10   19.122 < 0.0000000000000002 ***
Price.Progresso  -2457.29      19.24 -127.685 < 0.0000000000000002 ***
SouthTRUE         -676.93      16.27  -41.617 < 0.0000000000000002 ***
WestTRUE           -49.64      17.95   -2.766              0.00568 **
EastTRUE          1181.18      18.26   64.698 < 0.0000000000000002 ***
High_Income        360.05      14.90   24.165 < 0.0000000000000002 ***
Low_Income        -291.40      14.74  -19.763 < 0.0000000000000002 ***
Season             836.22      12.46   67.103 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1674 on 88399 degrees of freedom
Multiple R-squared:  0.3945,    Adjusted R-squared:  0.3945
```

- The formula would be: Yhat (predicted sales)= 915.70(Price.Campbell) + 613.87(Price.PL)+-2457.29(Price.Progresso) + -676.93(South) + -49.64(West) + 1181.18 (East) + -291.40(Low_Income) + 360.05(High_Income) + 836.22(Season) + 3086.50

- *Explain the results of the regression model (model strength, variable importance, relationship between the predictor and dependent variables). Use 1st tab in file.*
  - I came about this model by first looking at the variables that could possibly be correlated with sales of Progresso Soup. I combined them in a linear model, which yielded a moderately powerful R squared of .39. To make sure I wasn't missing out on any other variable combinations, I used bidirectional stepwise selection. Ultimately, it suggested the combination that I had originally proposed. Since the adjusted R-squared is also .39, it's safe to say that all the variables added to the model contribute a statistically significant predicative power (shown by the T values and low P values) that is not random.
  - The clearest relationship between the IVs and DV is with soup prices. As the price of Progresso soup goes up, it has a negative relationship with sales of Progresso. Conversely, the prices of the other types of soup have a positive relationship. As you would expect, people buy less of Progresso soup when it is more expensive and more when it is cheaper.
  - If estimating sales for a particular store or area, or planning advertising allocation, the region variables also showed powerful, yet quite intuitive influence. Naturally, the model predicts that stores in the South (where it is warmer) will sell less soup, while stores in the East (where it is cooler) will sell more. The West, with its varying climate, was less powerful and less significant.
  - Season also proved to be a powerful predictor of sales, as one might expect. However, the power of this predictor was diluted by the fact that seasonal differences in weather are not as stark in areas of the South, MidWest and West. If you limit the data to only the Eastern region, you see that the absolute effect of season is more than double.
- **Part 3 – Montana State Case**
  - For the Montana State University case, determine the outcome of the test on the homepage. How did each wording option perform? What do you recommend the University do to improve the performance of the webpage? Provide your answer and reasoning in just a couple of sentences.

- All tests were significant, except for Test C, with an alpha of .05. One of the significant tests, Test A, yielded a lower value than the control group.
- I would recommend going with Test B, Connect, as it had the largest percentage difference, aka Lift.

Full code: Part 1

```
library(stringr)
library(corrplot)
library(caTools)
library(car)
library(pscl)
censusData <- read.csv("adult.csv", strip.white=TRUE)



censusData$workclass <- gsub("?", NA, censusData$workclass, fixed=TRUE)
censusData$workclass <- as.factor(censusData$workclass)

ageMean <- mean(censusData$age)
ageSD <- sd(censusData$age)
z <- (censusData$age - ageMean)/ageSD
z3 <- subset(censusData, z >= 3)
z2 <- subset(censusData, z <2)
z1 <- subset(censusData, z <1)

hist(censusData$age)
par(mfrow=c(4,1))
hist((censusData$age), col="blue",border="white")
hist((z1$age), col="blue",border="white")
hist((z2$age), col="blue",border="white")
hist((z3$age), col="blue",border="white")
plot(censusData$workclass)

censusData$income <- NULL


for (i in 1:nrow(censusData)) {
  if (censusData$salary[i] == "<=50K") {
    censusData$income[i] <- 0
  }
  else {
    censusData$income[i]<- 1
  }
```

```
}

censusData$income <- factor(censusData$income)



censusData$relationship <- NULL
for (i in 1:nrow(censusData)) {
  if (censusData$marital.status[i] == "Married-spouse-absent" ||censusData$marital.status[i] == "Married-civ-
spouse" || censusData$marital.status[i] == "Married-AF-spouse") {
    censusData$relationship[i] <- 1
  }
  else {
    censusData$relationship[i] <- 0
  }
}




censusData["newSex"] <- NULL
for (i in 1:nrow(censusData)) {
  if (censusData$sex[i] == "Male") {
    censusData$newSex[i] <- 1

  }
  else  {
    censusData$newSex[i] <- 0

  }

}


censusData["capitalTotal"] <- censusData$capital.gain +  censusData$capital.loss

corrDF <- subset(censusData, select=c("age", "education.number", "hours.per.week", "income", "newSex",
"relationship", "capitalTotal"))



as.factor(censusData$occupation)
set.seed(101)
sample <- sample.split(censusData, SplitRatio=.75)
train = subset(censusData, sample == TRUE)
test = subset(censusData, sample == FALSE)
train <- na.omit(train)
test <- na.omit(test)
logitmodel <- glm(income ~ education.number + age + relationship + capitalTotal +hours.per.week + newSex,
data=train, family = "binomial")
summary(logitmodel)
vif(logitmodel)
```

```
fitted.results <- predict(logitmodel, newdata=subset(test,select=c("age", "hours.per.week", "education.number",
"income", "newSex", "relationship", "capitalTotal")), type='response')
fitted.results <- ifelse(fitted.results > 0.5, 1, 0)
misClasificError <- mean(fitted.results != test$income)
print(paste('Accuracy', 1-misClasificError))
```

```
nullmod <- glm(censusData$income~1, family="binomial")
mcfaddenR2 <- 1-logLik(logitmodel)/logLik(nullmod)
mcfaddenR2
summary(mcfaddenR2)
library(caret)
confusionMatrix(fitted.results, test$income)
library(ROCR)
p <- fitted.results
pr <- prediction(p, test$income)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)

auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

Full code: Part 2

```
soup <- read.csv("soup.csv")
library(psych)
library(MASS)
names(soup)
head(soup, lines=10)
summary(soup)
soup["Season"] <- NULL
for (i in 1:nrow(soup)) {
 if (soup$Month[i] >= 10 || soup$Month[i] <=2) {
  soup$Season[i] <- 1
 }
 else {
  soup$Season[i] <- 0
 }
}
soup$Region <- factor(soup$Region)
soup$MidWest<- as.logical(0)
soup$West<- as.logical(0)
soup$South<- as.logical(0)
soup$East<- as.logical(0)
for (i in 1:nrow(soup)) {
 if (soup$Region[i] == "MidWest") {
  soup$Midwest[i] <- as.logical(1)
 }
```

```
 else if (soup$Region[i] =="West") {
  soup$West[i] <- as.logical(1)
 }
 else if (soup$Region[i] =="East") {
  soup$East[i] <- as.logical(1)
 }
 else {
  soup$South[i] <- as.logical(1)
 }
}
shareWinter <- (sum(soup$Sales.Progresso[soup$Season == 1]) / sum(soup$Category_Sales[soup$Season == 1])) *
100
shareNonWinter <- (sum(soup$Sales.Progresso[soup$Season == 0]) / sum(soup$Category_Sales[soup$Season ==
0])) * 100
shareNonWinter
shareWinter
options(scipen=999)
fit <-
lm(soup$Sales.Progresso~Price.Campbell+Price.PL+Price.Progresso+South+West+East+MidWest+High_Income+Lo
w_Income+Season, data=soup)
summary(fit)
step <- stepAIC(fit, direction="both")
step$anova
summary(step)
soupEast <- subset(soup, Region=="East")
fitEast <-
lm(soupEast$Sales.Progresso~Price.Campbell+Price.PL+Price.Progresso+South+West+East+MidWest+High_Income
+Low_Income+Season, data=soupEast)
fitEast
summary(fitEast)
step <- stepAIC(fitEast, direction="both")
step$anova
summary(step)
```